

AD-A161 341

EFFICIENCY LOSS WITH THE KAPLAN-MEIER ESTIMATOR(U)

1/1

FLORIDA STATE UNIV TALLAHASSEE DEPT OF STATISTICS

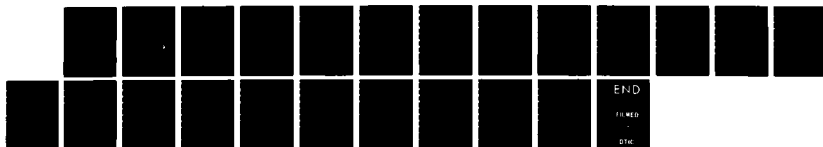
M HOLLANDER ET AL AUG 85 FSU-STATISTICS-M707

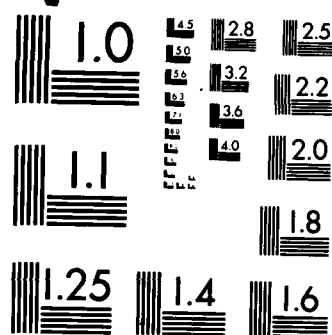
UNCLASSIFIED

AFOSR-TR-85-0975 F49620-85-C-0007

F/G 12/1

NL





MICROCOPY RESOLUTION TEST CHART  
NATIONAL BUREAU OF STANDARDS-1963-A



AD-A161 341

Efficiency Loss  
with the Kaplan-Meier Estimator

by

Myles Hollander, Frank Proschan, and James Sconing

FSU Statistics Report-M707, TR-  
AFOSR Technical Report No. 85-181

August, 1985

The Florida State University  
Department of Statistics  
Tallahassee, Florida 32306-3033

NOV 21 1985  
S  
A

Research sponsored by the Air Force Office of Scientific Research, AFSC, USAF, under Grant AFOSR 85-C-0007. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

1981 AMS Subject Classification: Primary 62G20; Secondary 62N05

Key Words and Phrases: Censored model, Kaplan-Meier estimator, Proportional hazards.

85 11 15\_037

DTIC FILE COPY

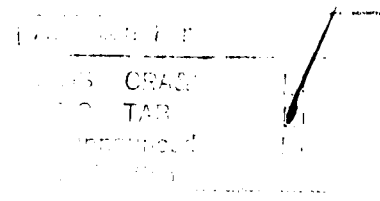
Efficiency Loss  
with the Kaplan-Meier Estimator

by

Myles Hollander, Frank Proschan, and James Sconing

ABSTRACT

We consider the proportional hazards model where the distribution  $G$  of the censoring random variable is related to the distribution  $F$  of the lifetime random variable via  $(1 - G) = (1 - F)^\beta$ . Nonparametric estimators of  $F$  are developed for the case where  $\beta$  is unknown and the case where  $\beta$  is known. Of interest in their own right, these estimators also enable us to study the robustness of the Kaplan-Meier estimator (KME) in a nonparametric model for which it is not the preferred estimator. Comparisons are based on asymptotic efficiencies and exact mean square errors. We also compare the KME to the empirical survival function, thereby providing, in a nonparametric setting, a measure of the loss in efficiency due to the presence of censoring.



A1

# 1. INTRODUCTION

In the usual censorship model we wish to estimate a life distribution  $F(x) = P(X \leq x)$  when lifelengths  $X_1, X_2, \dots, X_n$ , independent and identically distributed (i.i.d.) from  $F$  are under censorship by  $Y_1, \dots, Y_n$ , i.i.d. from censoring distribution  $G$ .  $X_i$  and  $Y_i$  are mutually independent for  $i = 1, \dots, n$  and  $F$  and  $G$  are continuous with densities  $f$  and  $g$  which are strictly positive on  $[0, \infty)$ . The actual observations consist of  $(Z_i, \delta_i)$ ,  $i = 1, \dots, n$ , where  $Z_i = \min(X_i, Y_i)$  and  $\delta_i = I(X_i \leq Y_i)$  where  $I(A)$  is the indicator function of the set  $A$ .

As an estimator of the survival function  $S(t) = 1 - F(t)$ , the Kaplan-Meier (1958) estimator (KME) has received considerable attention. It is defined as

$$\hat{S}_K(t) = \prod_{Z_{(i)} \leq t} c_{in}^{\delta_{(i)}} I(Z_{(n)} \geq t), \quad t \in (0, \infty), \quad (1.1)$$

where  $c_{in} = (n-i)(n-i+1)^{-1}$ ,  $Z_{(1)} < \dots < Z_{(n)}$  are the ordered  $Z_i$ 's, and  $\delta_{(i)}$  is the  $\delta$  corresponding to  $Z_{(i)}$ . The product over an empty set is defined to be zero. Some authors (cf. Wellner 1985) use a slightly different version of the KME defined by

$$\tilde{S}_K(t) = \prod_{Z_{(i)} \leq t} c_{in}^{\delta_{(i)}}, \quad t \in (0, \infty). \quad (1.2)$$

Equation (1.2) differs from (1.1) on  $[Z_{(n)}, \infty)$  if  $\delta_{(n)} = 0$ . While (1.1) is always zero on  $[Z_{(n)}, \infty)$ , (1.2) is strictly positive there if  $\delta_{(n)} = 0$  and thus in some samples  $\tilde{S}_K$  is not a true distribution function.

The KME has been studied in great detail. Weak convergence has been studied by Efron (1967), Breslow and Crowley (1974), Meier (1975), Gill (1983), and Wellner (1985). Strong consistency was established by Peterson (1977) and

Langberg, Proschan, and Quinzi (1980). Optimality properties were established by Wellner (1982). Small-sample properties have been studied by Chen, Hollander, and Langberg (1982), and Wellner (1985). Most of the properties developed in these papers require only minimal assumptions (e.g., continuity of  $F$  and  $G$ ). The KME is also the generalized maximum likelihood estimator. These properties along with the ease of computation, ease of interpretation, and easily estimated asymptotic variance (Greenwood's formula) have made the KME standard for estimating  $S(t)$ .

Miller (1983) terms the KME "seductive" in that it is very tempting to use. He studies the KME's efficiency loss when compared to the maximum likelihood estimate (MLE) in parametric models. Emoto (1984) compares the KME with parametric MLE's on the basis of mean square error. She considers both the case when the parametric model is correctly specified and the case when it is misspecified. Not surprisingly the KME performs poorly compared to MLE's in a fully parametric setting. For example for  $F$  and  $G$  exponential, Miller (1983) shows that the asymptotic efficiency of the KME with respect to the MLE tends to zero as  $t \rightarrow 0$  and as  $t \rightarrow \infty$ .

We study the properties of the KME by considering the proportional hazards model which lies between the parametric model and the fully nonparametric model. The proportional hazards model is nonparametric in the sense that  $F$  is unknown, but it possesses more structure than the fully nonparametric model assumed for the KME. By considering the proportional hazards model we can see how well the KME performs in a setting for which it is not optimal, thus investigating its robustness. Furthermore, our efficiency results in conjunction with those of Miller (1983) and Emoto (1984) allow us to determine the degree to which the KME efficiency losses are due to (1) full parametrization of the distribution of  $X$  and  $Y$  and (2) the presence of additional structure governing  $X$  and  $Y$ .

The proportional hazards model is:

Definition 1.1.  $(X, Y)$  follows a proportional hazards model if for some  $\beta > 0$ ,

$$1 - G(t) = \{S(t)\}^\beta, \quad t \in (0, \infty). \quad (1.3)$$

Expression (1.3) is equivalent to

$$R_G(t) = \beta R_F(t), \quad t \in (0, \infty), \quad (1.4)$$

where  $R_F(t) = -\log S(t)$ ,  $R_G(t) = -\log(1 - G(t))$ , the cumulative hazard functions of  $F$  and  $G$  respectively.

Proportional hazards has been used in censored models in the past. Efron (1967) uses the special case of exponential random variables to compare efficiencies for various two-sample tests. Koziol and Green (1976) derive a Cramér-von Mises statistic for testing a goodness-of-fit hypothesis that  $F = F_0$ ,  $F_0$  completely specified. Csörgő and Horváth (1984) improved upon the Koziol-Green test in that Koziol and Green required that  $\beta$  be known whereas Csörgő and Horváth do not need this assumption. Chen, Hollander, and Langberg (1982) and Wellner (1985) use proportional hazards to compute moments of the KME. Chen, Hollander, and Langberg use the form of the KME listed in (1.1) while Wellner uses (1.2).

In Section 2 we develop an estimator  $\hat{S}_p$  (2.3) for estimating  $S$  in the proportional hazards model when  $\beta$  is unknown. We compare  $\hat{S}_p$  with the KME in terms of asymptotic efficiency, exact bias and exact mean square error. In Section 3 we advocate the maximum likelihood estimator  $\tilde{S}_p$  (3.1) of  $S$  in the case of proportional hazards when  $\beta$  is known. One efficiency result is that the asymptotic efficiency of the KME with respect to the MLE is  $(\beta + 1)^{-1}$ . Since  $(\beta + 1)^{-1}$  is equal to  $P(X < Y)$ , this is a readily interpretable measure of efficiency. In

Section 4 the KME is compared with the empirical survival function. This comparison provides a measure of the efficiency loss due to censoring.

## 2. PROPORTIONAL HAZARDS, PROPORTIONALITY CONSTANT UNKNOWN

Assume that the proportional hazards model is known to hold. Then the KME is no longer the generalized maximum likelihood estimator. The extra information that  $1 - G(p) = \{S(p)\}^\beta$  should be utilized. Let  $T_n = n^{-1} \sum_{i=1}^n \delta_i$ . Then  $T$  is asymptotically normal with mean  $(\beta + 1)^{-1} = P(X < Y)$  and asymptotic variance:

$$AV(n^{1/2}T_n) = \beta\{(\beta + 1)^{-2}\}. \quad (2.1)$$

Let  $H(t)$  be the survival distribution for  $Z$ . Then  $H(t) = \{S(t)\}^{\beta+1}$ . Let  $\hat{H}_n(t) = n^{-1} \sum_{i=1}^n I(Z_i > t)$ , the usual empirical survival estimator for  $H(t)$ . Then  $A(t) = n^{1/2}\{\hat{H}_n(t) - H(t)\}$  converges weakly to a Gaussian process with mean 0 and covariance structure, for  $s \leq t$ , given by

$$\text{Cov}\{A(s), A(t)\} = \{1 - H(s)\}H(t) \text{ for } 0 < s \leq t < \infty. \quad (2.2)$$

Now our goal is to estimate  $S(t) = \{H(t)\}^{1/(\beta+1)}$ . A natural choice is

$$\hat{S}_p(t) = \{\hat{H}_n(t)\}^{T_n} \text{ for } t \in (0, \infty). \quad (2.3)$$

We use the following result of Allen (1963).

**Theorem 2.1.** The pair  $(X, Y)$ ,  $0 < P(X < Y) < 1$ , follows the proportional hazards model if and only if the random variables  $Z = \min(X, Y)$  and  $\delta = I(X \leq Y)$  are independent.

From Theorem 2.1 it follows that the random vectors  $(Z_1, \dots, Z_n)$  and  $(\delta_1, \dots, \delta_n)$  are independent under the proportional hazards model. Thus the



statistics  $T_n$  and  $\hat{H}_n(t)$  are independent. This, together with (2.1), (2.2) and the fact that  $g(x, y) = x^y$  has first partial derivatives which are continuous at  $[(1 - H(t))H(t), \beta(\beta + 1)^{-2}]$  imply that  $\hat{S}_p(t)$  converges weakly to a Gaussian process (cf. Serfling pg. 124) with mean 0 and asymptotic variance given by, for  $t \in (0, \infty)$ ,

$$AV\{n^{1/2}\hat{S}_p(t)\} = (\beta + 1)^{-2}\{H(t)\}^{(1-\beta)(\beta+1)^{-1}}\{1 - H(t)\} + \beta\{(\beta + 1)^{-2}\{\log H(t)\}^2\{H(t)\}^{2/(\beta+1)}\} \quad (2.4)$$

or, equivalently,

$$AV\{n^{1/2}\hat{S}_p(t)\} = (\beta + 1)^{-2}\{S(t)\}^{1-\beta}[1 - \{S(t)\}^{\beta+1}] + \beta\{\log S(t)\}^2\{S(t)\}^2. \quad (2.5)$$

It is interesting to note that the asymptotic variance of  $\hat{S}_p$  may decrease as  $\beta$  increases. This is not true of the asymptotic variance of the KME. From (2.5), we find

$$\frac{d}{d\beta} [AV\{n^{1/2}\hat{S}_p(t)\}] = -2(\beta + 1)^{-3}[\{S(t)\}^{1-\beta} - \{S(t)\}^2] - (\beta + 1)^{-2}\{S(t)\}^{1-\beta}\log S(t) + \{S(t)\log S(t)\}^2. \quad (2.6)$$

For  $\beta$  in a neighborhood of 1 and  $t$  close to 0, the right-hand side of (2.6) is less than zero. It seems counterintuitive that an estimator should improve as censoring increases. However, note that when  $\beta$  is close to 1, the distribution of  $Y$  is almost the same as that of  $X$ . Consequently observing  $Y$  is almost as informative as observing  $X$ . Thus this result is not surprising after all.

Note that the estimator in (2.3) jumps at both the observed  $X$ 's and the observed  $Y$ 's. Ebrahimi (1984) proposed an estimator in the proportional hazards model which jumps only at the observed  $X$ 's. Also note that the estimator in

(2.3) drops to zero after  $Z_{(n)}$  with one exception. In the case where  $T_n = 0$ ,  $\hat{S}_p(t) \equiv 1$ . In this pathological case our estimate for  $\beta$  is infinite.

$\hat{S}_p$  is also strongly consistent. Note that  $\hat{H}_n(t) \xrightarrow{a.s.} H(t)$  and  $T_n \xrightarrow{a.s.} T$  by the strong law of large numbers. Since  $g(x, y) = x^y$  is a continuous function,

$$\hat{S}_p(t) \xrightarrow{a.s.} \{H(t)\}^{1/(1+\beta)} = S(t) \{[1 - G(t)]^{1/(\beta+1)} \{S(t)\}^{-\beta/(\beta+1)}\}.$$

If the proportional hazards model holds then the term  $\phi(t) = \{[1 - G(t)]^{1/(\beta+1)} \{S(t)\}^{-\beta/(\beta+1)}\}$  reduces to 1. If the proportional hazards model does not hold then the term  $\phi(t)$  is a contaminating factor. The error in the estimator then depends on how far  $\phi(t)$  diverges from 1.

From (2.4) it is seen that the asymptotic variance can be estimated by

$$\hat{AV}\{n^{1/2}\hat{S}_p(t)\} = T^2 \left(\frac{n-i}{n}\right)^{1-2T} \left(\frac{i}{n}\right) + T(1-T) \{\log(\frac{n-i}{n})\}^2 \left(\frac{n-i}{n}\right)^{2T},$$

for  $Z_{(i)} \leq t < Z_{(i+1)}$ . This holds only for  $t < Z_{(n)}$ . Note also that if  $\beta = 0$ , (2.4) reduces to  $S(t)\{1 - S(t)\}$ , the asymptotic variance of the usual empirical survival function.

To compare  $\hat{S}_p$  with the KME, the asymptotic variance for the KME under the proportional hazards model must be computed. The estimator  $\hat{S}_K(t)$  is asymptotically normal with asymptotic variance (cf. Miller, 1981):

$$AV\{n^{1/2}\hat{S}_K(t)\} = \{S(t)\}^2 \int_0^t \frac{dF(u)}{\{S(u)\}^2 \{1 - G(u)\}}. \quad (2.7)$$

If  $1 - G = S^\beta$  then (2.7) reduces to

$$AV\{n^{1/2}\hat{S}_K(t)\} = (\beta + 1)^{-1} \{S(t)\}^2 [\{S(t)\}^{-(\beta+1)} - 1]. \quad (2.8)$$

The ratio of (2.5) to (2.8) is then

$$\alpha_1(t) \stackrel{\text{def}}{=} e(\hat{S}_K, \hat{S}_p) = (\beta + 1)^{-1} + \beta(\beta + 1) \{\log S(t)\}^2 [\{S(t)\}^{-(\beta+1)} - 1]^{-1}. \quad (2.9)$$

Theorem 2.2. The function  $\alpha_1(t)$  has the following properties:

- i)  $\lim_{t \rightarrow 0} \alpha_1(t) = (\beta + 1)^{-1}$ .
- ii)  $\lim_{t \rightarrow \infty} \alpha_1(t) = (\beta + 1)^{-1}$ .
- iii)  $\alpha_1(t) \geq (\beta + 1)^{-1}, 0 < t < \infty$ .

Proof: (i) Use L'Hospital's rule on the second term in (2.9) to obtain

$$\lim_{t \rightarrow 0} \alpha_1(t) = (\beta + 1)^{-1} + \lim_{t \rightarrow 0} -2\beta \{\log S(t)\} \{S(t)\}^{(\beta + 1)} = (\beta + 1)^{-1}.$$

(ii) Use L'Hospital's rule twice on the second term in (2.9) to obtain

$$\lim_{t \rightarrow \infty} \alpha_1(t) = (\beta + 1)^{-1} + \lim_{t \rightarrow \infty} -2\beta \{(\beta + 1)^{-1}\} \{S(t)\}^{(\beta + 1)} = (\beta + 1)^{-1}.$$

(iii) Note that the second term in (2.9) is always positive. ||

Table 1 gives some values for  $\alpha_1(t)$  for X exponential with parameter 1 and Y exponential with parameter  $\beta$ . Note that the values for  $\alpha_1(t)$  initially increase and then decrease. The value of t for which this change occurs is given by the solution to the equation  $(\beta + 1)t = 2[1 - \exp\{-t(\beta + 1)\}]$ . Table 1 also suggests that  $\alpha_1(t)$  decreases as  $\beta$  increases.  $\beta$  increasing is equivalent to censoring increasing stochastically. Thus Table 1 suggests that the efficiency of the KME with respect to  $\hat{S}_p$  decreases as censoring increases stochastically. We have been unable to prove this.

While Table 1 gives values for X and Y exponential, these values hold for any proportional hazards model. Consider the random variables R(X) and R(Y), where R(.) is the cumulative hazard function. Then R(X) and R(Y) are exponential random variables with parameters 1 and  $\beta$  respectively. To find the efficiency

of the KME with respect to  $\hat{S}_p(t)$  for this case, compute  $R(t)$  and use Table 1 with  $R(t)$  in place of  $t$ .

Finite sample comparisons can also be made using the method of Chen, Hollander, and Langberg (1982). These authors compute bias and variance for the KME under the proportional hazards model. Wellner (1985) does the same using (1.2) rather than (1.1). These methods can also be applied to  $\hat{S}_p$ . This gives

$$E\{\hat{S}_p(t)\}^\alpha = \sum_{j=0}^n \sum_{k=0}^n \binom{n}{j} \binom{n}{k} \left(\frac{n-j}{n}\right)^{\alpha k/n} \{S(t)\}^{(n-j)(\beta+1)} [1-\{S(t)\}^{\beta+1}]^j \cdot \{(\beta+1)^{-1}\}^j \{\beta/(\beta+1)\}^{n-j}. \quad (2.10)$$

We use (2.10) to calculate bias and mean square error for  $\hat{S}_p(t)$  when  $X$  and  $Y$  are exponential with parameters 1 and  $\beta$  respectively. Table 2 gives numerical values for bias and mean square error for  $\hat{S}_K(t)$ ,  $\tilde{S}_K(t)$ , and  $\hat{S}_p(t)$ . The values for  $\hat{S}_K$  and  $\tilde{S}_K$  are obtained from Wellner (1985). From Table 2 we see that  $\hat{S}_p$  is biased high; in fact, its bias exceeds that of both KME's. This is perhaps due to the pathological case  $T=0$ . The mean square error however is typically smaller than that of the KME, particularly when  $\beta = t = 2.0$ . The cases for which the mean square error of the KME is smaller seem to correspond to the cases for which the bias of  $\hat{S}_p$  is large compared to that of the KME. The mean square error and bias for  $\hat{S}_p$  tend to increase in  $\beta$  and decrease in  $n$ . (An exception occurs in the bias values for  $\beta = t = 2.0$ .) The values for the general proportional hazards case can be obtained, as previously seen, by considering exponential variables with  $R(t)$ , the hazard rate, taking the place of  $t$ .

### 3. PROPORTIONAL HAZARDS, PROPORTIONALITY CONSTANT KNOWN

Suppose the proportional hazards model is known to hold with  $\beta$  known. In this case an estimator analogous to  $\hat{S}_p$  is:

$$\tilde{S}_p(t) = \{\hat{H}_n(t)\}^\gamma \text{ for } t \in (0, \infty). \quad (3.1)$$

where  $\gamma = (\beta + 1)^{-1}$  and  $\hat{H}_n(t)$  is the empirical estimator for the  $Z_i$ 's.

It follows (Zehna, 1966) that  $\tilde{S}_p(t)$  is the maximum likelihood estimator for  $S(t)$ . Further, analogous to Section 2, if the model is correctly specified,  $\tilde{S}_p(t)$  is strongly consistent:

$$\tilde{S}_p(t) \xrightarrow{a.s.} S(t) \{[1 - G(t)]^{1/(\beta+1)} \{S(t)\}^{-\beta/(\beta+1)}\} = S(t).$$

If the proportional hazards model does not hold or if  $\beta$  is misspecified, then  $\tilde{S}_p(t)$  will not converge to  $S(t)$  and the error depends on how much the term  $\{[1 - G(t)]^{1/(\beta+1)} \{S(t)\}^{-\beta/(\beta+1)}\}$  differs from 1.

The estimator  $\tilde{S}_p$  converges weakly to a Gaussian process with mean  $S(t)$  and asymptotic variance given by:

$$AV\{n^{1/2}\tilde{S}_p(t)\} = (\beta + 1)^{-2} \{H(t)\}^{(1-\beta)/(\beta+1)} \{1 - H(t)\} \quad (3.2)$$

or, equivalently,

$$AV\{n^{1/2}\tilde{S}_p(t)\} = (\beta + 1)^{-2} \{S(t)\}^{1-\beta} [1 - \{S(t)\}^{\beta+1}]. \quad (3.3)$$

From (3.2) the asymptotic variance can be estimated by

$$\hat{AV}\{n^{1/2}\tilde{S}_p(t)\} = (\beta + 1)^{-2} \left(\frac{n-i}{n}\right)^{1-\beta} \left(\frac{i}{n}\right),$$

for  $Z_{(i)} \leq t < Z_{(i+1)}$ ,  $i = 1, \dots, n-1$ . Again the estimator jumps at both failure times and censoring times. To compare  $\tilde{S}_p$  with the KME, compute the ratio of

(3.3) to (2.8). This yields:

$$\alpha_2(t) \stackrel{\text{def}}{=} e(\hat{S}_K, \tilde{S}_P) = (\beta + 1)^{-1}, \text{ independent of } t.$$

Note that  $e(\hat{S}_K, \hat{S}_P)$  decreases as  $\beta$  increases. Recall from Theorem 2.2 that  $(\beta + 1)^{-1}$  is also the value of  $\alpha_1(t)$  at both extremes of  $t$ . Note that  $(\beta + 1)^{-1} = P(X < Y)$  and this represents the proportion of values for which a failure occurs. Recall that  $\tilde{S}_P$  jumps at both the observed failure times and the observed censoring times while  $\hat{S}_K$  jumps only at the observed failure times,  $n \cdot P(X < Y)$  in expectation.

As in Section 2, exact finite sample results can be obtained. Analogous calculations yield

$$E\{\tilde{S}_P(t)\}^\alpha = \sum_{j=0}^n \binom{n}{j} \left(\frac{n-j}{n}\right)^{\alpha/(\beta+1)} \{S(t)\}^{(n-j)(\beta+1)} [1 - \{S(t)\}^{\beta+1}]^j. \quad (3.4)$$

Bias and mean square error are calculated from (3.4). Table 3 gives the values for the case  $X$  and  $Y$  exponential with parameters 1 and  $\beta$  respectively. The biases for the  $\beta$  known case are higher than for the  $\beta$  unknown case. The mean square errors are everywhere smaller, sometimes half as small as those for which  $\beta$  is unknown. Note that  $\tilde{S}_K(t)$  has the smallest mean square error when  $t = 1.0$  and  $\beta = 2.0$ . However when  $t = 2.0$  and  $\beta = 2.0$ ,  $\tilde{S}_K(t)$  does substantially worse than each of the other competitors with mean square error six times as great as that of  $\tilde{S}_P$ . The mean square error and bias of  $\tilde{S}_P$  decrease with  $n$  and increase with  $\beta$ .

#### 4. LOSS IN EFFICIENCY DUE TO CENSORING

When there is no censoring, the KME reduces to the empirical survival function, the latter being the estimator of choice in the fully nonparametric non-censored model. Thus by comparing the KME to the empirical survival function, we obtain, in a nonparametric context, a measure of efficiency loss due to the presence of censoring.

From (2.7) the asymptotic variance of  $\hat{S}_E(t)$  is given by

$$AV\{n^{1/2}\hat{S}_E(t)\} = S(t)\{1 - S(t)\}. \quad (4.1)$$

The ratio of (2.7) to (4.1) is then

$$\alpha_3(t) \stackrel{\text{def}}{=} e(\hat{S}_E, \hat{S}_K) = S(t)\{1 - S(t)\}^{-1} \int_0^t \frac{dF(u)}{\{S(u)\}^2\{1 - G(u)\}}. \quad (4.2)$$

$\alpha_3(t)$  has the following interpretation. Roughly speaking,  $\hat{S}_K$  requires  $n \cdot e(\hat{S}_E, \hat{S}_K)$  observations in the censored model to do as well as  $\hat{S}_E$  does with  $n$  observations from the non-censored model.

##### Theorem 4.1.

- (i)  $\lim_{t \rightarrow 0} \alpha_3(t) = 1.$
- (ii)  $\lim_{t \rightarrow 0} \alpha_3(t) = \infty.$
- (iii)  $\alpha_3(t)$  is increasing in  $t.$
- (iv)  $\alpha_3(t)$  increases as censoring increases stochastically.

Proof: (i) We have

$$\lim_{t \rightarrow 0} \alpha_3(t) = \lim_{t \rightarrow 0} S(t)\{F(t)\}^{-1} \int_0^t \frac{dF(u)}{\{S(u)\}^2\{1 - G(u)\}} = \lim_{t \rightarrow 0} \{F(t)\}^{-1} \int_0^t \frac{dF(u)}{\{S(u)\}^2\{1 - G(u)\}}.$$

Using L'Hospital's rule,

$$\lim_{t \rightarrow 0} \alpha_3(t) = \lim_{t \rightarrow 0} \frac{f(t)}{\{1 - G(t)\} \{S(t)\}^2 f(t)} = \lim_{t \rightarrow 0} [\{1 - G(t)\} \{S(t)\}^2]^{-1} = 1.$$

(ii) Let  $\epsilon > 0$  be given and choose  $t_1$  such that  $1 - G(t_1) < \epsilon$  for  $t > t_1$ . Choose  $t_2$  such that  $S(t_1) - S(t_2) > (1/2)S(t_1)$ . Then

$$\begin{aligned} \lim_{t \rightarrow \infty} \alpha_3(t) &\geq S(t_2) \{F(t_2)\} \int_0^{t_2} \frac{dF(u)}{\{S(u)\}^2 \{1 - G(u)\}} \\ &\geq S(t_2) \{F(t_2)\}^{-1} \int_{t_1}^{t_2} \frac{dF(u)}{\{S(u)\}^2 \{1 - G(u)\}} \geq (2/\epsilon) S(t_2) \int_{t_1}^{t_2} \{S(u)\}^{-2} \{-dS(u)\} \\ &\geq \{2S(t_2)/\epsilon\} [\{S(u)\}^{-1} \Big|_{t_1}^{t_2}] = \{2S(t_2)/\epsilon\} [\{S(t_2)\}^{-1} - \{S(t_1)\}^{-1}] \geq \epsilon^{-1}. \end{aligned}$$

(iii)  $\frac{d}{dt} \{\alpha_3(t)\} =$

$$\begin{aligned} &S(t) f(t) [F(t) \{1 - G(t)\} \{S(t)\}^2]^{-1} + \int_0^t \frac{dF(u)}{\{S(u)\}^2 \{1 - G(u)\}} [f(t) \{F(t)\}^{-2}] \\ &= f(t) \{F(t)\}^{-1} [\{S(t) \{1 - G(t)\}\}^{-1} - \{F(t)\}^{-1} \int_0^t \frac{dF(u)}{\{S(u)\}^2 \{1 - G(u)\}}], \end{aligned}$$

which is positive if

$$\{F(t)\}^{-1} \{1 - G(t)\} S(t) \int_0^t \frac{dF(u)}{\{S(u)\}^2 \{1 - G(u)\}} \leq 1. \quad (4.3)$$

The right-hand side of (4.3) is less than

$$\{F(t)\}^{-1} S(t) \int_0^t \{S(u)\}^{-2} dS(u) = \{F(t)\}^{-1} S(t) [\{S(u)\}^{-1} \Big|_0^t] = 1.$$

(iv) Note that if censoring increases stochastically,  $1 - G(t)$  decreases for every value of  $t$ . This implies that  $\alpha_3(t)$  is increasing. ||



These results indicate that when  $t$  is small, censoring is not very critical, but as  $t$  increases the censoring has more influence. Consequently for functionals of  $S(t)$  which involve large values of  $t$ , the KME must be used with caution.

Acknowledgement: We gratefully acknowledge Edsel Peña for checking the efficiency expressions and the bias and mean square error calculations.

REFERENCES.

- Allen, W. R. (1963), "A Note on Conditional Probability of Failure When Hazards are Proportional," Operations Research, 11, 658-659.
- Breslow, N., and Crowley, J. (1974), "A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship," Annals of Statistics, 2, 437-453.
- Chen, Y. Y., Hollander, M., and Langberg, N. A. (1982), "Small-Sample Results for the Kaplan-Meier Estimator," Journal of the American Statistical Association, 77, 141-144.
- Csörgö, S. and Horváth, L. (1981), "On the Koziol-Green Model for Random Censorship," Biometrika, 68, 391-401.
- Ebrahimi, N. (1984), "Nonparametric Estimation of Survival Functions for Incomplete Observations," Unpublished Manuscript, Department of Statistics, Northern Illinois University.
- Efron, B. (1967), "The Two-Sample Problem With Censored Data," Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 4, 831-852.
- Emoto, S. (1984), "The Preferred Choice Between the Maximum Likelihood Estimator and the Kaplan-Meier Estimator," Stanford University Department of Biostatistics Technical Report No. 96.
- Gill, R. D. (1983), "Large Sample Behaviour of the Product-Limit Estimator on the Whole Line," Annals of Statistics, 11, 49-58.
- Kaplan, E. L., and Meier, P. (1958), "Nonparametric Estimation from Incomplete Data," Journal of the American Statistical Association, 53, 457-481.
- Koziol, J. A., and Green, S. B. (1976), "A Cramér-von Mises Statistic for Randomly Censored Data," Biometrika, 63, 465-474.
- Langberg, N. A., Proschan, F., and Quinzi, A. T. (1981), "Estimating Dependent Life Lengths with Applications to the Theory of Competing Risks," Annals of Statistics, 9, 152-167.
- Meier, P. (1975), "Estimation of a Distribution Function from Incomplete Observations," in Perspectives in Probability and Statistics, Ed. J. Gani, 67-87, Sheffield, England: Applied Probability Trust.
- Miller, R. G. (1981), Survival Analysis. New York: John Wiley & Sons.
- Miller, R. G. (1983), "What Price Kaplan-Meier?", Biometrics, 39, 1077-1082.

- Peterson, A. V. (1977), "Expressing the Kaplan-Meier Estimator as a Function of Empirical Subsurvival Functions," Journal of the American Statistical Association, 72, 854-858.
- Serfling, R. J. (1980), Approximation Theorems of Mathematical Statistics. New York: John Wiley & Sons.
- Wellner, J. A. (1982), "Asymptotic Optimality of the Product Limit Estimator," Annals of Statistics, 10, 595-602.
- Wellner, J. A. (1985), "A Heavy Censoring Limit Theorem for the Product Limit Estimator," Annals of Statistics, 13, 150-162.
- Zehna, P. W. (1966), "Invariance of Maximum Likelihood Estimators," Annals of Mathematical Statistics, 37, 744.

Table 1. Asymptotic Efficiency of  $\hat{S}_K$  with Respect  
to  $\hat{S}_P$  under Proportional Hazards with  $\beta$  Unknown.

$t \backslash \beta$	.1	.2	1/3	.5	1.0	2.0
.1	.9186	.8522	.7812	.7130	.5903	.5048
.2	.9270	.8687	.8082	.7524	.6627	.6253
.3	.9344	.8832	.8313	.7854	.7189	.7033
.4	.9409	.8957	.8509	.8126	.7611	.7471
.5	.9466	.9063	.8672	.8345	.7910	.7642
.6	.9515	.9153	.8806	.8516	.8103	.7611
.7	.9556	.9227	.8911	.8645	.8208	.7436
.8	.9590	.9286	.8993	.8736	.8238	.7164
.9	.9618	.9333	.9052	.8793	.8208	.6835
1.0	.9640	.9368	.9091	.8821	.8130	.6477
1.1	.9656	.9392	.9113	.8824	.8016	.6114
1.2	.9668	.9406	.9119	.8805	.7873	.5760
1.3	.9676	.9412	.9112	.8769	.7712	.5428
1.4	.9679	.9411	.9093	.8718	.7538	.5124
1.5	.9679	.9403	.9065	.8655	.7358	.4850
1.6	.9676	.9389	.9028	.8582	.7176	.4608
1.7	.9670	.9370	.8985	.8502	.6996	.4397
1.8	.9662	.9347	.8937	.8417	.6820	.4215
1.9	.9651	.9320	.8884	.8329	.6652	.4061
2.0	.9639	.9291	.8828	.8239	.6492	.3930

Table 2. Exact Values of Bias and Mean Square Error under Proportional Hazards, with  $\beta$  Unknown.

t	$\beta$	n	-Bias $\{\hat{S}_K(t)\}$	-Bias $\{\tilde{S}_K(t)\}$	-Bias $\{\hat{S}_P(t)\}$	MSE $\{\hat{S}_K(t)\}$	MSE $\{\tilde{S}_K(t)\}$	MSE $\{\hat{S}_P(t)\}$
.5	.5	10	.0001	.0000	.0040	.0218	.0280	.0235
		15	.0000	.0000	.0026	.0185	.0185	.0155
		20	.0000	.0000	.0019	.0138	.0138	.0115
		25	.0000	.0000	.0015	.0111	.0111	.0092
		30	.0000	.0000	.0013	.0092	.0092	.0077
.5	1.0	10	.0017	.0001	.0077	.0318	.0335	.0279
		15	.0001	.0000	.0041	.0291	.0218	.0174
		20	.0000	.0000	.0029	.0162	.0162	.0128
		25	.0000	.0000	.0023	.0129	.0129	.0102
		30	.0000	.0000	.0019	.0107	.0107	.0085
.5	2.0	10	.0250	.0022	.0323	.0646	.0478	.0548
		15	.0063	.0004	.0131	.0359	.0313	.0290
		20	.0017	.0001	.0068	.0250	.0229	.0189
		25	.0004	.0000	.0044	.0184	.0181	.0142
		30	.0001	.0000	.0034	.0150	.0149	.0115
1.0	.5	10	.0054	.0010	.0111	.0362	.0334	.0314
		15	.0012	.0002	.0053	.0226	.0219	.0197
		20	.0003	.0000	.0033	.0164	.0163	.0143
		25	.0001	.0000	.0024	.0130	.0129	.0113
		30	.0000	.0000	.0019	.0107	.0107	.0094
1.0	1.0	10	.0339	.0073	.0419	.0600	.0475	.0521
		15	.0141	.0022	.0217	.0380	.0318	.0320
		20	.0061	.0008	.0127	.0265	.0236	.0220
		25	.0027	.0003	.0082	.0200	.0186	.0164
		30	.0012	.0001	.0058	.0160	.0154	.0130

Table 2. (continued)

t	$\beta$	n	-Bias $\{\hat{S}_K(t)\}$	-Bias $\{\tilde{S}_K(t)\}$	-Bias $\{\hat{S}_P(t)\}$	MSE $\{\hat{S}_K(t)\}$	MSE $\{\tilde{S}_K(t)\}$	MSE $\{\hat{S}_P(t)\}$
1.0	2.0	10	.1549	.0486	.1505	.1092	.0803	.1043
		15	.1142	.0268	.1184	.0898	.0588	.0797
		20	.0814	.0160	.0906	.0714	.0464	.0627
		25	.0601	.0101	.0700	.0586	.0383	.0501
		30	.0447	.0065	.0548	.0485	.0324	.0405
2.0	.5	10	.0285	.0193	.0320	.0246	.0258	.0213
		15	.0188	.0095	.0229	.0182	.0170	.0154
		20	.0129	.0053	.0172	.0142	.0127	.0119
		25	.0090	.0031	.0133	.0114	.0101	.0095
		30	.0064	.0019	.0106	.0095	.0084	.0078
2.0	1.0	10	.0742	.0723	.0757	.0287	.0522	.0263
		15	.0623	.0472	.0656	.0255	.0349	.0222
		20	.0532	.0335	.0572	.0228	.0263	.0194
		25	.0459	.0249	.0505	.0206	.0218	.0172
		30	.0398	.0191	.0449	.0187	.0178	.0153
2.0	2.0	10	.1228	.2079	.1063	.0224	.1293	.0340
		15	.1190	.1674	.1174	.0225	.0952	.0231
		20	.1156	.1419	.1162	.0225	.0761	.0215
		25	.1125	.1238	.1136	.0223	.0639	.0210
		30	.1097	.1101	.1103	.0222	.0552	.0207

Table 3. Exact Values of Bias and Mean Square Error under Proportional Hazards with  $\beta$  Known.

t	$\beta$	n	-Bias $\{\tilde{S}_p(t)\}$	MSE $\{\tilde{S}_p(t)\}$	t	$\beta$	n	-Bias $\{\tilde{S}_p(t)\}$	MSE $\{\tilde{S}_p(t)\}$
.5	.5	10	.0084	.0196	1.0	2.0	10	.1709	.0887
		15	.0053	.0127			15	.1300	.0680
		20	.0039	.0094			20	.1010	.0529
		25	.0031	.0075			25	.0798	.0415
		30	.0026	.0062			30	.0640	.0330
.5	1.0	10	.0163	.0198	2.0	.5	10	.0370	.0174
		15	.0097	.0118			15	.0276	.0127
		20	.0070	.0085			20	.0215	.0098
		25	.0055	.0067			25	.0172	.0079
		30	.0045	.0055			30	.0141	.0065
.5	2.0	10	.0484	.0392	2.0	1.0	10	.0801	.0217
		15	.0240	.0176			15	.0694	.0188
		20	.0151	.0102			20	.0611	.0165
		25	.0110	.0071			25	.0543	.0147
		30	.0088	.0056			30	.0486	.0132
1.0	.5	10	.0196	.0255	2.0	2.0	10	.1239	.0205
		15	.0115	.0157			15	.1205	.0204
		20	.0080	.0113			20	.1174	.0201
		25	.0062	.0088			25	.1146	.0199
		30	.0051	.0073			30	.1120	.0196
1.0	1.0	10	.0551	.0405					
		15	.0325	.0239					
		20	.0215	.0158					
		25	.0155	.0114					
		30	.0121	.0087					

A161341

## SECURITY CLASSIFICATION OF THIS PAGE

## REPORT DOCUMENTATION PAGE

1. REPORT NUMBER FSU M707 AFOSR 85-181	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and subtitle)  Efficiency Loss with the Kaplan-Meier Estimator		5. TYPE OF REPORT & PERIOD COVERED Technical
6. AUTHOR(s)  Myles Hollander, Frank Proschan, & James Sconing		6. PERFORMING ORG. REPORT NUMBER
7. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Florida State University Tallahassee, Florida 32306-3033		8. CONTRACT OR GRANT NUMBER(s)  AFOSR F49620-85-C-0007
8. CONTROLLING OFFICE NAME AND ADDRESS The U.S. Air Force Air Force Office of Scientific Research Bolling Air Force Base, DC 20332		10. PROGRAM ELEMENT, PROJECT, TASK AREA, & WORK UNIT NUMBERS  G1102F 2304 A5
9. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		12. REPORT DATE August, 1985
		13. NUMBER OF PAGES 19
		15. SECURITY CLASS. (of this report)
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

10. DISTRIBUTION STATEMENT (of this report)  
distribution unlimited

11. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from report)

12. SUPPLEMENTARY NOTES

13. KEY WORDS

Censored model, Kaplan-Meier estimator, Proportional hazards

14. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We consider the proportional hazards model where the distribution  $G$  of the censoring random variable is related to the distribution  $F$  of the lifetime random variable via  $(1 - G) = (1 - F)^{\beta}$ . Nonparametric estimators of  $F$  are developed for the case where  $\beta$  is unknown and the case where  $\beta$  is known. Of interest in their own right, these estimators also enable us to study the robustness of the Kaplan-Meier estimator (KME) in a nonparametric model for which it is not the preferred estimator. Comparisons are based on asymptotic efficiencies and exact mean square errors. We also compare the KME to the empirical survival function, thereby providing, in a nonparametric setting, a measure of the loss in efficiency due to the presence of censoring.



**END**

**FILMED**

**1-86**

**DTIC**